基于 filter+wrapper 模式的特征选择算法 *

周传华 [1,2]，柳智才 [1†]，丁敬安 [1]，周家亿 [3]

(1. 安徽工业大学 管理科学与工程学院，安徽 马鞍山 243002; 2. 中国科学技术大学 计算机科学与技术学院，合肥 230026; 3. 早稻田大学 IPS 学院，日本 东京)

摘　要： 特征选择是数据挖掘、机器学习和模式识别中始终面临的一个重要问题。针对类和特征分布不均时，传统信息增益在特征选择中存在的选择偏好问题，提出了一种基于信息增益率与随机森林的特征选择算法。该算法结合 filter 和 wrapper 模式的优点，首先从信息相关性和分类能力两个方面对特征进行综合度量，然后采用序列前向选择（sequential forward selection，SFS）策略对特征进行选择，并以分类精度作为评价指标对特征子集进行度量，从而获取最优特征子集。实验结果表明，本文算法不仅能够达到特征空间降维的效果，而且能够有效提高分类算法的分类性能和查全率。

关键词： 信息增益率；随机森林；特征选择；filter 模式；wrapper 模式

中图分类号： TP301.6　　doi: 10.3969/j.issn.1001-3695.2018.01.0024

Feature selection algorithm based on Filter + Wrapper pattern

Zhou Chuanhua[1,2], Liu Zhicai[1], Ding Jing'an[1], Zhou Jiayi[3]

(1. *School of Management Science & Engineering Anhui University of Technology*, *Maanshan AnHui* 243002, *China*; 2. *School of Computer Science & Technology*, *University of Science &Technology of China*, *Hefei* 230026, *China*; 3. *Graduate School of Information*, *Production & Systems Waseda University*, *Tokyo*, *Japan*)

Abstract: Feature selection is one of the most important issues in data mining, machine learning and pattern recognition. Aiming at the problem of preference of traditional information gain algorithm in feature selection when the class and feature are unevenly distributed, this paper proposes a new feature selection algorithm based on information gain ratio and random forest. The proposed algorithm combined with the advantages of Filter and Wrapper modes. First, a comprehensive measurement of features is carried out from two aspects of information correlation and classification ability. Second, Sequential Forward Selection (SFS) strategy is used to select the features, and the classification accuracy is used as the evaluation index to measure the feature subset. Finally, obtain the optimal feature subset. The experimental results show that the proposed algorithm can not only achieve the effect of dimension reduction in feature space, but also effectively improve the classification performance and recall rate of classification algorithm.

Key words: information gain ratio; random forest; feature selection; filter mode; wrapper mode

0　引言

特征选择是指在保证特征集合分类性能的前提下，从一组原始特征集合中选出具有代表性的特征子集，以达到降低特征空间维数的过程[1]。特征选择作为数据预处理中的关键步骤，根据是否依赖机器学习算法，可以分为过滤式(filter)和封装式(wrapper)两种。过滤式特征选择算法利用数据的内在特性对选取的特征子集进行评价和选择，独立于机器学习算法，该类算法通常运行效率较高，但结果较差；而封装式特征选择算法则依赖于机器学习算法的分类精度作为特征子集选择的评价准则，该类算法效率较低，但选择的特征集合性能较优。

常见的特征选择算法有信息增益（information gain, IG）、粗糙集、神经网络、互信息[2]（mutual information, MI）和卡方统计等。其中，IG 是一种有效的特征选择算法，多用于文本分类中。文献[3-6]研究了传统 IG 特征选择算法在文本分类中的应用，发现在类和特征分布不均时，传统信息增益在特征选择

收稿日期：2018-01-16；修回日期：2018-03-09　基金项目：国家自然科学基金资助项目（71371013，71772002）；安徽省留学人员创新项目择优资助计划（2016）

作者简介：周传华（1965-），男，安徽马鞍山人，教授，博士，主要研究方向为机器学习、数据挖掘、智能算法研究；柳智才（1993-），男（通信作者），硕士研究生.主要研究方向为机器学习、模式识别、智能优化（lzc646211927@163.com）；丁敬安（1991-），男，硕士研究生，主要研究方向为数据挖掘、数据分析；周家亿（1993-），男，硕士研究生，主要研究方向为数据分析、智能算法、模式识别.

中性能下降问题，并从特征项的频数以及基于词频的类内分布和类间分布等角度提出改进。文献[7,8]中则通过分析传统 IG 和 CHI 算法的优缺点，并将两种算法进行结合提出一种组合算法。文献[9]中，罗养霞等人针对软件胎记特征选择问题，提出了一种基于层次聚类与信息度量的过滤式特征选择算法。该算法通过构建信息增益函数和惩罚函数，选择出具有高区分性和最小冗余的软件胎记特征。文献[10]中，尹建芹等人以 IG 为基础研究特征的分类能力与其支持度之间的关系，并证明了具有高支持度或低支持度的特征具有有限的分类能力，从而为频繁模式挖掘在分类问题中进行特征选择奠定了理论基础。文献[11]中，刘云等人针对用户网络行为进行属性推断问题，基于 IG 度量特征重要性，提出了两种面向概率性特征选择算法的改进策略，从而解决特征空间高维问题和提高算法效率。文献[9~11]中虽然都使用了 IG 来度量特征的重要性与分类能力从而进行特征选择，但是他们都没有考虑 IG 算法存在的选择偏好问题。

文献研究表明，传统的 IG 算法在特征选择方面虽然具有一定的有效性，但是当数据量较大、数据中类和特征分布不均时，其本身的选择偏好问题会突显出来，导致其性能急剧下降，而信息增益率(gain ratio, GR)算法能够通过添加惩罚因子降低选择偏好的发生。并且，一些特征选择算法仅强调了特征空间维度的降低，没有考虑到特征集合的分类性能。因此，本文提出了一种基于 filter+wrapper 模式特征选择算法 FSIGR(feature selection based on importance and gain rate)，该算法结合了 filter 和 wrapper 模式的优点，在保证选择的特征子集性能较优的前提下，最大限度的提高了本文算法的运行效率。FSIGR 算法主要分为两个阶段：过滤和封装。在过滤阶段，首先使用 GR 对特征与类别之间的信息相关性进行度量，并删除无关特征，从而有效降低特征空间维度，提高算法运行效率，然后使用随机森林基于特征分类能力对相关特征进行重要度测评，并从特征与类别之间的信息相关性和分类能力两个方面对特征进行综合度量。在封装阶段，首先在综合度量的基础上对特征进行排序，然后使用 SFS 策略对单个特征进行选择并使用分类器对特征子集分类性能进行评估，以达到特征空间降维和提高特征集合分类性能的效果，从而选出最优特征子集。

1 基础理论

1.1 熵与信息增益率

信息熵作为信息论中的基本概念，是用于度量随机变量不确定性的一种数学表达，也是对变量本身或变量集合所含有的平均信息量的一种度量，通常用 $H(X)$ 表示。设 $X=\{x_1,x_2,\ldots,x_m\}$ 与 $Y=\{y_1,y_2,\ldots,y_m\}$ 是两个随机变量，$p(x_i)$ 和 $p(y_i)$ 为概率密度函数。则随机变量 X 的熵 $H(X)$ 定义为

$$H(X)=-\sum_{i=1}^{m}p(x_i)log_2 p(x_i) \tag{1}$$

随机变量 X 和 Y 的条件熵定义为：

$$H(X|Y)=\sum_{i=1}^{m}p(y_i)H(X|Y=y_i) \tag{2}$$

条件熵 $H(X|Y)\le H(X)$ 用来衡量变量 X 和 Y 的相关性，若变量 X 和 Y 不相关，则 $H(X|Y)=H(X)$；若变量 X 和 Y 相关，则 $H(X|Y)<H(X)$，且 $H(X)-H(X|Y)$ 值越大，变量 X 和 Y 相关性越强。

信息增益是一种无量纲的度量标准，它是对两个随机变量之间相关信息量的度量，其值越大说明变量之间的相关性越强。信息增益具有非对称性，它能从非线性的角度对特征之间的相关性进行度量。信息增益与熵、条件熵的关系为

$$IG(X|Y)=H(X)-H(X|Y) \tag{3}$$

由式（3）可以看出 $IG(X|Y)$ 值越大，说明变量 X 和 Y 相关性越强。其中 $IG(X|Y)$ 表示变量 Y 的信息增益。

在信息系统中，经常使用信息增益来衡量某个特征对信息系统分类的贡献，来降低样例中噪声的敏感度。但由于信息增益存在偏好选择分支较多的特征，导致过拟合的发生。因此，在使用时经常引入惩罚因子，来对分支较多的特征进行惩罚，即信息增益率：

$$GR(X|Y)=\frac{IG(X|Y)}{H(Y)} \tag{4}$$

由式（4）可以看出，随机变量 Y 的信息增益率与其信息增益成正比，与其信息熵成反比。因此，当随机变量 Y 取值较多时，$GR(X|Y)$ 会随着 $H(Y)$ 的增大而减小，在一定程度上降低了选择偏好的发生。

1.2 随机森林与重要度测评

随机森林（random forest, RF）是一种集成学习算法，它使用随机重采样技术和节点随机分裂技术构建多棵决策树，并根据投票机制产生最后的结果。由于 RF 对于存在噪声和缺失值的数据具有很好的鲁棒性，并且具有较快的学习速度，其变量重要性度量可以作为高维数据的特征选择工具，因此近年来已经被广泛应用于各种分类、预测、特征选择以及异常点检测问题中[12~14]。

基于 RF 的特征重要度测评有两种度量方法,一种是基于袋外数据(out of bag, OOB)检测误差的方法，称为平均准确率降低(mean decrease accuracy, MDA)；另一种是基于 Gini 不纯度的方法，称为平均基尼系数降低（mean decrease gini, MDG）。两种方法都是通过判断特征对 RF 分类性能的影响来确定该特征的重要性，值下降的越多表示特征越重要。其中，MDA 是通过添加随机噪声和 OOB error 检测误差的方法来对特征进行度量，并确定特征的重要程度[15-16]。

算法主要步骤如下：

设随机森林包括 M 棵分类回归树。为测度第 j 个特征属性对输出变量的重要性，对随机森林中的每棵分类树进行处理。对第 $i(i=1,2,\ldots,M)$ 棵分类回归树：

a）计算第 i 棵分类回归树基于袋外观测的预测误差，记为 e_i。

b）随机打乱袋外观测在第 j 个特征属性上的取值顺序，重新建立第 i 棵分类回归树并袋外观测进行预测。

c）重新计算第 i 棵分类回归树的预测误差,记为 e_i^j 。$\varepsilon_i^j = e_i - e_i^j$ 为第 j 个特征属性添加噪声导致的第 i 棵分类回归树预测误差的变化。

重复上述步骤，最终得到 M 个预测误差的变化。$MDA^j = \frac{1}{M}\sum_{i=1}^{m}\varepsilon_i^j$ 即为第 j 个输人变量加噪声导致的随机森林总体预测误差的平均变化,它测度了第 j 个输人变量的重要性。

2 FSIGR 算法

FSIGR 算法主要分为两个部分：过滤阶段和封装阶段。在过滤阶段本文提出一种综合评估算法(comprehensive evaluation algorithm, CEA)。在封装阶段本文提出一种递归删除算法(recursive deletion algorithm, RDA)。首先采用 CEA 算法对特征进行过滤和综合性评估，以尝试从不同的维度增强对特征的度量，提高算法的运行效率。然后采用 RDA 算法对特征进行选择，可以在不牺牲算法精度的情况下降低特征的波动性，从而产生最优特征子集。设数据集为 $\boldsymbol{D}$，特征属性集为 F ={ $f_i \mid i = 1 \ldots v$ }，则 FSIGR 算法流程如下：

2.1 综合评估算法 CEA

首先计算每个特征关于类别特征的 GR，若其 GR 等于 0，则表示该特征和类别特征不相关,并从特征集合中删除该特征。然后对数据集中的特征分别使用 GR 和 MDA 算法从信息相关性和分类能力两个方面进行重要度度量，最后对度量结果分别进行标准化处理。具体公式如下：

$$\tilde{m}_i = \frac{m_i}{\sum_{i=1}^{v} m_i} \tag{5}$$

$$g_i = \frac{g_i}{\sum_{i=1}^{v} g_i} \tag{6}$$

其中：m_i 和 g_i 分别表示 MDA 和 GR 算法对特征 $f_i (i = 1 \ldots v)$的重要度度量值，$\tilde{m}_i$ 和 g_i 则分别表示其标准化后的值。并映射成权重向量 $\vec{\boldsymbol{c}_i} = \left(\tilde{m}_i, g_i\right)$，其中 $\tilde{m}_i$ 和 g_i 表示向量 $\vec{\boldsymbol{c}_i}$ 的坐标值。向量 $\vec{\boldsymbol{c}_i}$ 的长度则表示特征 f_i 的重要度。

最后根据 $\tilde{m}_i$ 和 g_i 值计算特征 f_i 的综合评估值 c_i 。

$$c_i = \sqrt{\left|\tilde{m}_i^2 + g_i^2\right|} \tag{7}$$

由式（7）可以看出，CEA 算法以 GR 和 MDA 为基础，通过将 $\tilde{m}_i$ 和 g_i 的值进行标准化和向量化对特征 f_i 进行综合度量，既考虑了特征 f_i 与类别特征之间相关性，又考虑到了特征 f_i 的分类能力，增强了对特征的度量，降低了特征的波动性。从而选择出最大相关和最大分类能力的特征，并删除冗余特征。与文献[8]中的 IG 相比，本文使用 GR 计算特征的信息相关性有效降低了 IG 的选择偏好问题；与文献[12]中 MDA+MDG 的方法相比，本文从特征信息相关性和分类能力两种不同的维度对特征进行综合度量，降低了特征的波动性。算法描述如下：

算法 1：CEA 算法

输入：数据集 $\boldsymbol{D}$，特征集合 F ={ $f_i \mid i = 1 \ldots v$ }。

过程：

分别计算特征 f_i 关于类别特征的 GR 值 g_i，若 g_i=0，则删除特征 f_i，$F = F - \{f_i\}$；

使用随机森林计算特征 f_i 重要度 MDA 值，并记为 m_i；

运用式（5）（6）分别对 m_i，g_i 进行标准化，得到 m_i，g_i；

根据式（7）计算特征 f_i 的综合评估值 c_i；

输出：特征综合评估值 c_i。

2.2 递归删除算法 RDA

SFS 算法描述：特征子集 F_i 从空集开始，每次选择一个特征 f_i 加入特征子集 F_i，使得特征函数 J(F)最优。SFS 算法是一种简单的贪心算法。

RDA 算法思想：根据综合评估值 c_i 对特征进行降序排序，然后运用 SFS 策略遍历特征空间，得到相应的特征集合 F_1、F_2、…、F_v，并使用分类器对该特征集合进行评估记为 a_i，若 $a_i < a_{i-1}$，则从集合 F 中删除 f_i 元素，记录当前最优特征子集 a_{temp} 并与全局最优特征子集 a_{max} 进行比较，若 $a_{max} < a_{temp}$，则 $a_{max} = a_{temp}$，重复上述操作，直至循环结束；若全局特征子集分类性能较优，a_{max} 不变，重复上述操作，直至循环结束，输出最优特征子集。

RDA 算法在综合评估的基础上，使用分类精度对每个特征子集的分类性能进行评估，可以在不牺牲算法精度的情况下降低特征子集的波动性，并删除重要度较小的冗余特征。每次遍历仅删除一个特征，并产生新的特征组合，扩大特征子集搜索空间的覆盖范围，从而选出最小冗余、性能最优的特征子集。与文献[8,12]中的简单过滤式方法相比，本文采用的过滤+封装模式，提高了特征子集的分类性能。算法描述如下：

算法 2：RDA 算法

输入：数据集 $\boldsymbol{D}$，特征集合 F ={ $f_i \mid i = 1 \ldots v$ }, c_i,$a_{max} = 0, F_{best} = \varnothing$。

过程：

1 根据特征 f_i 的综合测评度 c_i，对特征进行降序排序；

2 repeat

3 使用分类器对特征子集进行评估：首先对排序后的特征子集采用 SFS 搜索策略产生相应的特征子集 F_i，然后分别计算分类器在该特征子集 F_i 上的精确度 a_i，其中 i 表示特征子集中元素的个数；

4 flag = false

5 for a_i （$i = 1 \ldots v$）do

6 if $a_i < a_{i-1}$ then

7 flag = true

8 从集合 F 中删除特征 f_i，并记录删除特征 f_i 后分类器的精度为 a_{temp}；

9 if $a_{max} < a_{temp}$ then

10 $a_{max} = a_{temp}$，$F_{best} = F$

11 end if

```
12      break
13    end if
14   end for
15 until flag == false 达到终止条件
```

输出：最优特征子集 F_{best} 。

2.3 FSIGR 算法复杂度分析

本文算法的时间开销分为 CEA 算法和 RDA 算法两个部分。其中，时间开销主要体现在第二部分。根据文献[11]可知，若训练数据集的特征维数为 m，训练样本个数为 n，假设 RF 算法中基分类器的个数为 k，则 RF 算法的时间复杂度近似为 $O\left(kmn(\log n)^2\right)$，快速排序平均时间复杂度为 $\mathrm{O}(m(\log m))$。因此，CEA 算法的最大渐进时间复杂度为 $\mathrm{O}\left(3m+kmn(\log n)^2\right)$。RDA 算法中时间开销主要体现在行 2~8。其外层循环最多运行 m 次，相应的内层循环最多分别进行 $(m,m-1,m-2,\cdots,1)$ 次。因此，FSIGR 算法渐进最大时间复杂度可以表示为

$$\mathrm{O}\left(3m+kmn(\log n)^2\right)+\mathrm{O}(m(\log m))+\mathrm{O}\left(\frac{1}{2}*m(m-1)\right)$$

$$=\mathrm{O}(m\left(\frac{1}{2}*(m+5)+kn(logn)^2+logm\right)) \tag{8}$$

$$T(n)=\mathrm{O}\left(m^2\right) \tag{9}$$

由式（9）可以看出，FSIGR 算法的最大时间复杂度与特征维数近似平方，对高维数据具有较好的处理能力且具有很好的扩展性。由于本文算法在运行过程中临时占用存储空间大小与特征个数成线性正比关系。所以，空间复杂度可以表示为

$$S(n)=\mathrm{O}(m) \tag{10}$$

与 GR、MDA 以及 CFS 和文献[8]中的算法相比，本文算法由于是 Filter + Wrapper 模式，因此具有较高的时间复杂度，但是本文算法空间复杂度相对较低且算法性能较优。与 WFS 算法相比本文算法总体复杂度较低且算法性能较优，具有较好的实用性与扩展性。

3 实验及结果分析

为了验证 FSIGR 算法的有效性，本文将选用 CFS[17](Correlation-based Feature Selection) 算法、WFS[18](Wrapper Feature Selection)算法以及文献[8]中的算法与 FSIGR 算法进行实验对比。实验软硬件环境如下：操作系统为 Windows 10，CPU 为 Intel® Core™ i5-6300HQ @ 2.3 GHz，实验内存为 8 GB，主要实验平台为 WEKA[19]，语言为 Java。

3.1 实验数据

为了使实验中选取的数据集有广泛的代表性，从 UCI 数据集中选取了 4 个数据集进行测试，数据集描述如表 1 所示。这些数据集在分类数、实例数和特征维数方面均具有不同的特点，并且数据类型包含了数值型、标称型和混合型，同时有些数据集含有较多的缺失数据，可以较好地验证特征选择算法在实际数据集上的性能。

表 1　实验数据集

数据集	Breast Cancer	glass	credit	phishing
分类数	2	6	2	2
实例数	699	214	1000	11055
特征数	9	9	20	30
特征类型	nominal	numeric	nominal/numeric	nominal
是否有缺失值	YES	NO	NO	NO

3.2 实验方案

为了较好验证本文算法的有效性，本文进行两组对比实验：

实验 1　在每个数据集上，分别运用 4 种不同的特征选择算法进行特征选择，然后使用 Weka 中的 RF 算法对特征选择后的数据集进行训练，并采用 10 折交叉的方法进行验证，记录、对比实验结果。

实验 2　为了进一步验证 FSIGR 算法对分类结果的影响，首先使用不同的特征选择算法分别对 phishing 数据集进行特征选择，然后对选择后的数据集分别对 C4.5、KNN、RF 和 REPTree 分类模型进行训练，并采用 10 折交叉的方法进行验证，记录、对比实验结果。

3.3 评判指标

为了方便实验对比，本文采用精确度(accuracy)、召回率(recall)和 F-Measure 作为实验的评价指标，计算公式如下：

a)精确度：
$$accuracy=\frac{TP+TN}{TP+TN+FP+FN} \tag{11}$$

b)召回率：
$$recall=\frac{TP}{TP+FN} \tag{12}$$

c)F-measure:
$$F=\frac{2*accuracy*recall}{accuracy+recall} \tag{13}$$

其中：TP（true positive）：被正确分类为正例的样本数；FP（false positive）：被错误分类为正例的样本数；TN（true negative）：被正确分类为反例的样本数；FN（false negative）：被错误分类为反例的样本数。

3.4 结果分析

实验中，CFS 算法和 WFS 算法分别采用最佳优先(best first, BF)和贪婪算法(greedy stepwise, GS)两种搜索策略对特征进行选择；FSIGR 算法则采用 SFS 策略对特征进行选择。具体实验结果如下：

实验 1　图 1 中对 GR、MDA 和 FSIGR 三种特征选择算法的性能进行研究。图中纵坐标表示 RF 分类模型的分类精度。在实验中，根据文献[8]设置阈值为 0.01 对 GR 和 MDA 排序后的特征进行选择。由图 1 可以看出，在 4 个数据集中使用 FSIGR 算法产生的特征子集对 RF 分类模型进行训练分类精度最高，明显优于 GR 和 MDA 方法。因为在 FSIGR 算法 Filter 阶段通过将 GR 和 MDA 的值进行标准化和向量化对特征 f_i 进行综合

度量，既考虑了特征 f_i 与类别特征之间相关性，又考虑到了特征 f_i 的分类能力，增强了对特征的度量，降低了特征的波动性。从而选择出最大相关和最大分类能力的特征。在 Wrapper 阶段又以分类精度为评价指标选取分类性能最优的特征子集，因此其性能优于 GR 和 MDA 算法。

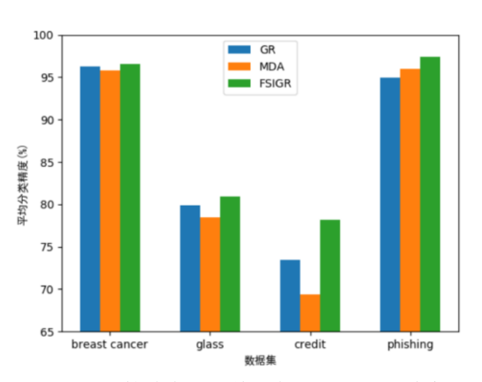


图 1　RF 算法在不同特征集合上的分类精度

表 2 中列出了 CFS、WFS、文献[8]和 FSIGR 算法在不同实验数据集上的性能比较，其中 SF 表示选出的特征集合的大小，Acc 表示 RF 算法在该特征集合上的算法精度，“—”表示该算法在相应数据集上没有进行实验。

由表 2 可以看出，RF 分类模型基于 FSIGR 算法在 4 个数据集分类精度分别为 0.966, 0.809, 0.782, 0.974，均优于 CFS，WFS 和文献[8]算法在 4 个数据集上的表现，证明了本文 FSIGR 算法具有较优的分类性能；在降维性能方面，FSIGR 算法在 glass 数据集上优于其他算法，但在 Breast Cancer, credit 和 phishing 数据集上，略低于其他算法。因为本文 FSIGR 算法采用 Filter＋Wrapper 模式进行特征选择，并从全局与局部对特征子集进行评价选择分类性能较优的特征子集。所以，在此过程中会牺牲一定的降维性能。结果表明，本文提出的算法在不同类型的数据集上均有较好的表现，能在提高特征集合分类性能的情况下对数据进行降维，具有鲁棒性。

表 2　不同特征选择算法的性能比较

特征选择算法		数据集			
		Breast Cancer	glass	credit	phishing
CFS	SF	9	8	3	9
	Acc	0.964	0.799	0.702	0.948
WFS	SF	2	7	3	29
	Acc	0.947	0.776	0.74	0.973
文献[8]算法	SF	——	——	——	17
	Acc	——	——	——	0.968
FSIGR	SF	**7**	**7**	14	23
	Acc	**0.966**	**0.809**	**0.782**	**0.974**

实验 2　表 3 中列出了在 phishing 数据集上使用不同特征选择算法在不同分类分类器下的特征选择结果。由表中数据可以看出，CFS 和文献[8]算法为作为 Filter 模式特征选择算法其特征选择结果与分类器无关且降维性能较优。Wrapper 模式的 WFS 特征选择算法和本文 Filter + Wrapper 模式的 FSIGR 算法降维性能稍弱，但由实验 1 的结果可知，本文 WFS 和 FSIGR 算法具有较优的分类性能且 FSIGR 算法均优于其他算法。

表 3　特征选择结果

特征选择算法	特征子集
WFS(BF)	$F=\{f_i \mid i=1\ldots30, i\neq5,9,11,12,16,18,19,22,23\}$
WFS(GS)	$F=\{f_i \mid i=1\ldots30, i\neq5,9,11,12,16,18,19,21,22,23,30\}$
FSIGR	$F=\{f_i \mid i=1\ldots30, i\neq4,10,11,19,20,23,28,30\}$
WFS(BF)	$F=\{f_i \mid i=1\ldots30, i\neq10\}$
WFS(GS)	$F=\{f_i \mid i=1\ldots30, i\neq4,10,18,19,21,23,30\}$
FSIGR	$F=\{f_i \mid i=1\ldots30, i\neq11,19,21,23,30\}$
WFS(BF)	$F=\{f_i \mid i=1\ldots30, i\neq4,5\}$
WFS(GS)	$F=\{f_i \mid i=1\ldots30, i\neq3\}$
FSIGR	$F=\{f_i \mid i=1\ldots30, i\neq4,11,18,19,21,23,30\}$
WFS(BF)	$F=\{f_i \mid i=1\ldots30, i\neq4,9,10,16,18,19,21,23,27,30\}$
WFS(GS)	$F=\{f_i \mid i=1\ldots30, i\neq4,5,11,16,21\}$
FSIGR	$F=\{f_i \mid i=1\ldots30, i\neq4,11,30\}$
CFS	$F=\{f_i \mid i=6,7,8,13,14,15,6,26,28\}$
文献[8]算法	$F=\{f_i \mid i=1,2,6,7,8,9,13,14,15,16,24,25,26,27,28,29,30\}$

注：F 表示特征子集，f_i 表示特征子集中的元素，i 为该元素在源数据集中的下标。

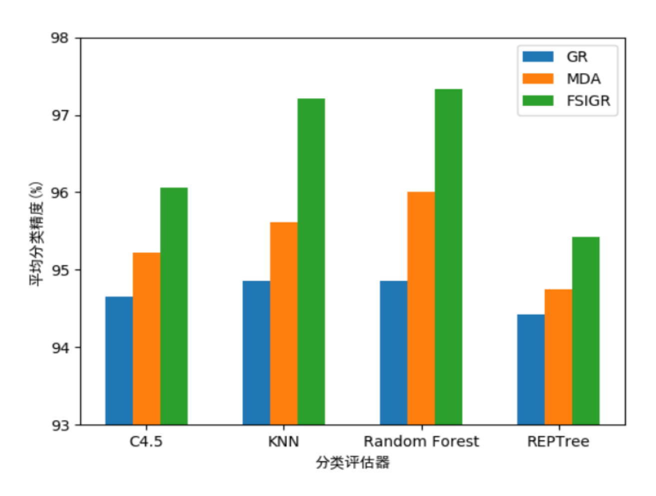


图 2　C4.5、KNN、RF 和 REPTree 四种不同的分类算法基于 GR、MDA 和 FSIGR 特征选择算法的精确度

图 2 中对比了在 C4.5、KNN、RF 和 REPTree 四种不同的分类模型下 GR、MDA 和 FSIGR 三种特征选择算法的性能。图中横坐标表示不同的分类模型，纵坐标表示相应分类模型的分类精度。从图 2 中可以看出，C4.5、KNN、RF 和 REPTree 四种分类模型在 FSIGR 算法上的分类精度最高。这是因为 FSIGR 特征选择算法能够从信息相关性和分类能力两个方面对特征进行综合度量，从而选出相关性强、冗余度低的最优特征子集，提高了分类模型的分类精度。本实验证明了 FSIGR 特征选择算法能有效降低特征子集的维度选出关键特征，从而提高分类模型的准确率。

图 3 中研究了在 C4.5、KNN、RF 和 REPTree 四种不同的分类模型下 CFS、WFS、文献[8]以及 FSIGR 四种特征选择算法性能。图中折线表示同一种特征选择算法基于不同分类模型选择的特征子集分类精度变化趋势。

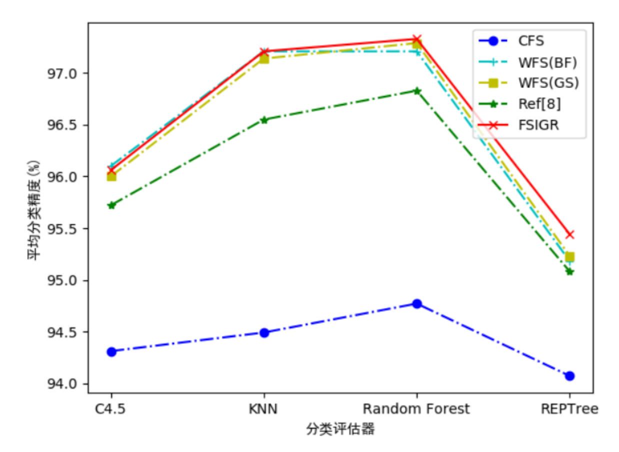


图 3 C4.5、KNN、RF 和 REPTree 四种不同的分类算法基于 WFS、FSIGR 等特征选择算法的精确度

由图 3 可以看出，基于 C4.5、KNN、RF 和 REPTree 四种不同的基分类模型，本文 FSIGR 算法均具有较优的表现，其分类性能明显优于 CFS 和文献[8]算法。其中，在 C4.5 和 KNN 分类模型上 FSIGR 与 WFS 算法表现相似，在 RF 和 REPTree 分类模型上 FSIGR 算法性能明显优于 WFS 等算法。因为与 CFS 和文献[8]算法相比，后两者为 Filter 模式特征选择算法，直接根据特征之间或者特征与类别之间的相应关系对特征进行选择，并未考虑特征子集整体的分类性能。而 FSIGR 算法结合 Filter 和 Wrapper 模式从单个特征与特征子集两个方面对特征子集进行选择，从而选出相关性强、冗余度低和分类能力较优的特征子集。与 WFS 算法相比，本文算法首先使用 Filter 模式对特征进行筛选，然后再以分类精度为指标对特征子集进行选择，这样可以首先排除部分冗余特征，然后在进行特征选择，降低了部分时间开销，所以本文 FSIGR 算法综合性能较优。

通过图 3 中可以发现，RF 在 4 种特征选择算法上的分类性能均优于 C4.5、KNN 和 REPTree 算法，那是因为 RF 为集成学习算法，它能够通过综合不同基分类器的分类结果增强集成学习算法的容错性和泛化能力，从而达到提高分类精度，分类召回率降低分类误差的目的。因此在 FSIGR 算法的 Filter 阶段，采用了 RF 算法的 MDA 对特征的分类能力进行测评，增强了对特征的度量，降低了特征的波动性。

表 4、5 描述实验 2 详细结果。其中，表 4 从平均绝对误差和召回率两个方面对 4 种特征选择算法的实验结果进行对比。由表 4 可以看出，本文算法的综合性能均优于其他算法，基于本文算法的分类模型具有较低的预测误差和较高的查全率。表 5 从 F-measure 和 AUC（area under ROC curve）两个方面对 4 种特征选择算法的实验结果进行对比。由表 5 可以看出，基于本文算法的分类模型其 F-Measure 和 AUC 值较大，分类能力较强，即本文算法选择的特征子集较优。

表 4 基于 WFS、FSIGR 等特征选择算法的实验结果 1

特征选择算法	搜索算法	C4.5		KNN		Random Forest		REPTree	
		平均绝对误差	recall	平均绝对误差	recall	平均绝对误差	recall	平均绝对误差	recall
CFS	(BF/GS)	0.087	0.943	0.075	0.945	0.075	0.948	0.087	0.941
WFS	BF	0.060	0.961	0.032	0.972	0.050	0.972	0.068	0.952
	GS	0.060	0.960	0.033	0.971	0.051	0.973	0.067	0.952
文献[8]算法	BF	0.061	0.957	0.039	0.965	0.049	0.968	0.067	0.951
FSIGR	SFS	**0.057**	**0.961**	**0.033**	**0.972**	**0.048**	**0.973**	**0.064**	**0.954**

表 5 基于 WFS、FSIGR 等特征选择算法的实验结果 2

特征选择算法	搜索算法	C4.5		KNN		Random Forest		REPTree	
		F-Socre	AUC	F-Socre	AUC	F-Socre	AUC	F-Socre	AUC
CFS	(BF/GS)	0.943	0.979	0.945	0.988	0.948	0.988	0.941	0.983
WFS	BF	0.961	0.984	0.972	0.990	0.972	0.996	0.952	0.985
	GS	0.960	0.984	0.971	0.990	0.973	0.996	0.952	0.983
文献[8]算法	BF	0.957	0.983	0.965	0.989	0.968	0.995	0.951	0.984
FSIGR	SFS	0.960	**0.985**	**0.972**	**0.990**	**0.973**	**0.996**	**0.954**	0.984

4 结束语

本文提出了一种 filter + wrapper 模式的特征选择算法 FSIGR 算法，在过滤阶段该算法首先以 GR 为度量标准对特征进行选择，从而选择出相关性强的特征，然后基于 GR 算法和 RF 算法从信息相关性和分类能力两个方面对特征综合度量。在封装阶段，根据综合度量对特征进行重新排序，并采用序列前向搜索策略和分类模型的分类精度作为评价标准寻找最优特征子集。通过 FSIGR 算法和 GR、MDA 算法的对比实验可以看出，FSIGR 算法的性能明显优于两种基本的算法，证明了 FSIGR

算法的有效性。通过 FSIGR 算法和 CFS、WFS 以及文献[8]算法的对比实验结果表明，FSIGR 算法在特征空间降维和提高分类精度方面均有较好的表现，并且 FSIGR 算法的性能明显优于文献[8]中的算法和 CFS 算法。通过对 FSIGR 算法从时间和空间两个方面进行复杂度分析发现，FSIGR 算法对高维数据有较好的处理能力，具有较好的实用性和扩展性。基于以上叙述，证明了本文 FSIGR 算法能够在保证特征子集分类性能的前提下，达到特征空间降维的效果，具有一定的有效性和实用性。

由于 FSIGR 算法中仅考虑到不同特征的重要性度量，而较少的考虑特征之间的相关性。因此，在 FSIGR 算法中考虑两两特征之间的相关性，是下一步工作的重点。

参考文献：

[1] Guyon I, Elisseeff A. An introduction to variable and feature selection [J]. Journal of Machine Learning Research, 2003, 3 (6): 1157-1182.

[2] Hoque N, Bhattacharyya D K, Kalita J K. MIFS-ND: a mutual information-based feature selection method [J]. Expert Systems with Applications, 2014, 41 (14): 6371–6385.

[3] Lee C, Lee G G. Information gain and divergence-based feature selection for machine learning-based text categorization [J]. Information Processing & Management, 2006, 42 (1): 155-165.

[4] 朱颢东，钟勇．基于改进的 ID3 信息增益的特征选择方法 [J]. 计算机工程, 2010, 36 (8): 37-39.

[5] 刘庆和，梁正友．一种基于信息增益的特征优化选择算法 [J]. 计算机工程与应用, 2011, 47 (12): 130-132.

[6] Wu G, Xu J. Optimized approach of feature selection based on information gain [J]. Computer Engineering & Applications, 2011, 47 (12): 157-161.

[7] 王光，邱云飞，史庆伟，等．集合 CHI 与 IG 的特征选择算法 [J]. 计算机应用研究, 2012, 29 (7): 2454-2456.

[8] Rajab K D. New hybrid features selection method: a case study on websites phishing [J]. Security & Communication Networks, 2017, 2017 (2): 1-10.

[9] 罗养霞，房鼎益．基于聚类分析的软件胎记特征选择 [J]. 电子学报, 2013, 41 (12): 2334-2338.

[10] 尹建芹，田国会，魏军，等．特征的支持度与其分类能力的关系研究 [J]. 电子学报, 2015, 43 (2): 248-254.

[11] 刘云，孙宇清，李明珠．面向社会化媒体用户评论行为的属性推断 [J]. 计算机学报, 2017: 1-15

[12] Hideko K, Hiroaki Y. Rapid feature selection based on random forests for high-dimensional data [J]. Ipsj Sig Notes, 2012, 2012: 1-7.

[13] 姚登举，杨静，詹晓娟．基于随机森林的特征选择算法 [J]. 吉林大学学报：工学版, 2014, 44 (1): 137-141.

[14] Han H, Guo X, Yu H. Variable selection using mean decrease accuracy and mean decrease gini based on random forest [C]// Proc of IEEE International Conference on Software Engineering and Service Science. 2017: 219-224.

[15] Wang H, Lin C, Peng Y, *et al.* Application of improved random forest variables importance measure to traditional Chinese chronic gastritis diagnosis [C]// Proc of IEEE International Symposium on It in Medicine and Education. 2008: 84-89.

[16] Nicodemus K K. Letter to the editor: on the stability and ranking of predictors from random forest variable importance measures [J]. Briefings in Bioinformatics, 2011, 12 (4): 369.

[17] Hall M A. Correlation-based feature selection for machine learning [D]. Hamilton, NewZealand: Waikato University 1999.

[18] Kohavi R, John G H. Wrappers for feature subset selection [J]. Artificial Intelligence, 1997, 97 (1-2): 273-324.

[19] Witten I H, Frank E. Data mining: practical machine learning tools and techniques with Java implementations [M]. 北京 机械工业出版社, 2012.